3/5/1 (Item 1 from file: 351)
DIALOG(R)File 351:Derwent WPI

0010495610 - Drawing available
WPI ACC NO: 2001-096482/200111
XRPX Acc No: N2001-073308
Processor communication apparatus for computer network, has main memory in which erroneous logic page numbers and number of erroneous pages are stored in preset position
Patent Assignee: NEC CORP (NIDE)
Inventor: KANO T; KANOH Y
Patent Family (3 patents, 2 countries)

Patent Number	Kind	Date	Application Number	Kind	Date	Update	
JP 2000330960	A	20001130	JP 1999134960	A	19990514	200111	B
JP 3376956	B2	20030217	JP 1999134960	A	19990514	200316	E
US 6678722	B1	20040113	US 2000568593	A	20000511	200405	E

Priority Applications (no., kind, date): JP 1999134960 A 19990514

Patent Details

Number	Kind	Lan	Pg	Dwg	Filing Notes
JP 2000330960	A	JA	16	11	
JP 3376956	B2	JA	16		Previously issued patent JP 2000330960

Alerting Abstract JP A
NOVELTY - Reading address and write-in address of data for communication between processors (1) are designated in a logic address. The flag shows the generation of error during communication from the logic processor number of sending element of packet, which assigns communication ID to sending element. The erroneous logic page number and number of erroneous pages are stored in preset position in a main memory (4).
USE - For computer network.
ADVANTAGE - Since the erroneous page lists are stored in main memory, consideration about size of area for communication during programming is avoided.
DESCRIPTION OF DRAWINGS - The figure shows the block diagram of processor communication apparatus.
1 Processor
4 Main memory

Title Terms/Index Terms/Additional Words: PROCESSOR; COMMUNICATE; APPARATUS ; COMPUTER; NETWORK; MAIN; MEMORY; ERROR; LOGIC; PAGE; NUMBER; STORAGE; PRESET; POSITION
Class Codes
International Classification (Main): G06F-015/167, G06F-015/177
(Additional/Secondary): G06F-012/00, G06F-012/08, G06F-012/10, G06F-015/00
US Classification, Issued: 709212000, 712029000, 711156000, 711202000

File Segment: EPI;
DWPI Class: T01
Manual Codes (EPI/S-X): T01-H03A; T01-H08; T01-M02C1

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2000－330960
（P2000－330960A）
(43)公開日　平成12年11月30日(2000.11.30)

(51)Int.Cl.⁷	識別記号	ＦＩ		テーマコード（参考）
Ｇ０６Ｆ　15/177	６７６	Ｇ０６Ｆ　15/177	６７６Ａ	５Ｂ００５
12/08		12/08	Ｈ	５Ｂ０４５
12/10		12/10	Ｆ	

審査請求　有　　請求項の数10　ＯＬ　（全 16 頁）

(21)出願番号　特願平11－134960

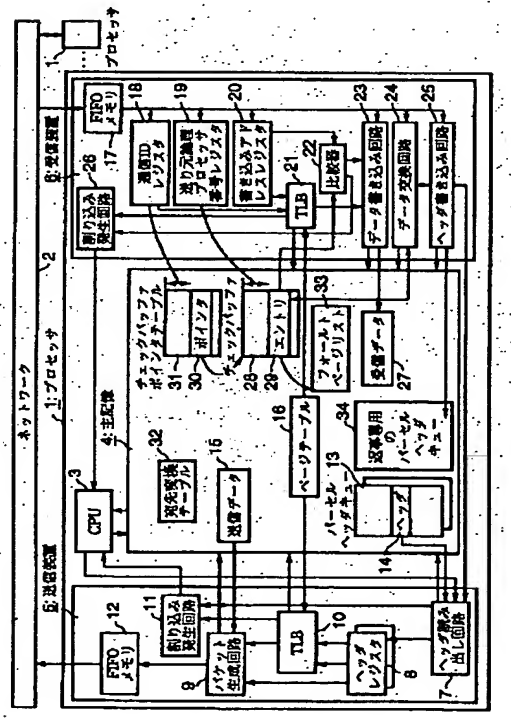
(22)出願日　平成11年５月14日(1999.5.14)

(71)出願人　000004237
日本電気株式会社
東京都港区芝五丁目７番１号
(72)発明者　加納　健
東京都港区芝五丁目７番１号　日本電気株式会社内
(74)代理人　100108578
弁理士　高橋　詔男　（外３名）

Ｆターム(参考)　5B005 JJ11 KK02 KK13 MM31 MM51 SS11 SS14
5B045 BB02 DD07 FF02 GG11

(54)【発明の名称】　プロセッサ間通信装置

(57)【要約】
【課題】　ページフォールトを回避するための処理によるプロセッサ間通信の性能の低下を回避し、かつ、プロセッサ間通信時のページフォールトに対応する。
【解決手段】　連続するデータのプロセッサ間通信の直後に、受信側でページフォールトが発生したかどうかを確かめるためのパケットを送る。ページフォールトが起こっていない場合には何もせず、ページフォールトが起こっていた場合には、ページフォールトした部分データだけを再送する。また、受信側で、ページフォールトしたページを記憶しておき、再送の時に使用する。さらに、ページフォールトが発生しても割り込みが頻繁に発生しないように、同じページでページフォールトしたかどうかを受信装置が確認し、同じならば割り込みを発生しない。



1

【特許請求の範囲】

【請求項１】　主記憶、送信装置、並びに受信装置を有するプロセッサが複数個、ネットワークによって結合された並列コンピュータに用いられるプロセッサ間通信装置であって、

プロセッサ間通信で通信されるデータの読み出しアドレスと書き込みアドレスが論理アドレスで指定され、

該送信装置と該受信装置内に論理アドレスから物理アドレスに変換する手段を備えると共に、

タスクごとに割り当てられた通信ＩＤと、パケットの送り元の論理プロセッサ番号から決定される主記憶上の場所に、該送り元の論理プロセッサから該通信ＩＤを割り当てられたタスクに送られたプロセッサ間通信時にページフォールトが発生したかどうかを示すフラグと、該プロセッサ間通信時に直近にページフォールトした論理ページ番号と、該プロセッサ間通信時にページフォールトしたページ数と、ページフォールトした論理ページ番号を格納する主記憶上の位置を記憶する手段を備えたプロセッサ間通信装置。

【請求項２】　前記受信装置内でのアドレス変換の結果、ページの状態がページアウトの場合に、

ページフォールトフラグ、該プロセッサ間通信時に直近にページフォールトした論理ページ番号、該プロセッサ間通信時にページフォールトしたページ数、ページフォールトした論理ページ番号を格納する主記憶上の位置を記憶する手段に該アドレス変換の結果に関するデータを記憶し、さらに該アドレス変換にかかる論理ページをページリストに加えることを特徴とする請求項１に記載のプロセッサ間通信装置。

【請求項３】　前記受信装置内でのアドレス変換の結果、ページの状態がページアウトの場合に、

前記アドレス変換にかかる論理ページを予め用意しておいたダミーページを割り当てることを特徴とする請求項２に記載のプロセッサ間通信装置。

【請求項４】　主記憶、送信装置、並びに受信装置を有するプロセッサが複数個、ネットワークによって結合された並列コンピュータに用いられるプロセッサ間通信装置であって、

プロセッサ間通信で通信されるデータの読み出しアドレスと書き込みアドレスが論理アドレスで指定され、

該送信装置と該受信装置内に論理アドレスから物理アドレスに変換する手段を備えると共に、

該論理アドレスを構成する論理ページの情報として、該論理ページの割り当てられた物理ページ番号と、ページの状態として下記（ｉ）～（ｉｖ）の状態を持つことを特徴とするプロセッサ間通信装置。

（ｉ）物理ページが割り当てられていない無効の状態；

（ｉｉ）現在主記憶になく外部記憶媒体に追い出されているページアウトの状態；

（ｉｉｉ）以前はページアウトの状態で現在ページイン

2

処理を行なっている際中であるページイン中の状態；

（ｉｖ）該主記憶上に割り付けられているページインの状態。

【請求項５】　主記憶、送信装置、並びに受信装置を有するプロセッサが複数個、ネットワークによって結合された並列コンピュータに用いられるプロセッサ間通信装置であって、

プロセッサ間通信で通信されるデータの読み出しアドレスと書き込みアドレスが論理アドレスで指定され、

該送信装置と該受信装置内に論理アドレスから物理アドレスに変換する手段を備えると共に、

該受信装置での論理アドレスから物理アドレスへの変換時に、タスクごとに割り当てられた通信ＩＤと、パケットの送り元の論理プロセッサ番号から決定される主記憶上の場所に記憶されている論理ページ番号と、変換しようとしている論理ページ番号とを比較する手段を持つプロセッサ間通信装置。

【請求項６】　前記受信装置内でのアドレス変換の結果、ページの状態がページイン中の場合には、

前記受信装置での論理アドレスから物理アドレスへの変換時に、タスクごとに割り当てられた通信ＩＤと、パケットの送り元の論理プロセッサ番号から決定される主記憶上の場所に記憶されている論理ページ番号と、変換しようとしている論理ページ番号とを比較し、

該論理ページ番号が同じ場合には何もせず、該論理ページ番号が異なった場合にだけ、その論理ページ番号をページリストに加え、直近にページフォールトした論理ページ番号として記憶し、ページフォールトしたページ数を１増やす、ことを特徴とする請求項５に記載のプロセッサ間通信装置。

【請求項７】　主記憶、送信装置、並びに受信装置を有するプロセッサが複数個、ネットワークによって結合された並列コンピュータに用いられるプロセッサ間通信装置であって、

プロセッサ間通信で通信されるデータの読み出しアドレスと書き込みアドレスが論理アドレスで指定され、

該送信装置と該受信装置内に論理アドレスから物理アドレスに変換する手段を備えると共に、

宛先プロセッサ内の主記憶上の、タスクごとに割り当てられた通信ＩＤと、パケットの送り元の論理プロセッサ番号から決定される場所のデータと、パケットに付加したデータとを交換し、自プロセッサの指定された主記憶上の位置に交換したデータを書き込む機能を持つパケットを持つプロセッサ間通信装置。

【請求項８】　プロセッサ間通信でデータを送った直後に、前記パケットを用いて、事前に送ったプロセッサ間通信の受信側でのページフォールトがあったかどうかを知ることを特徴とする請求項７に記載のプロセッサ間通信装置。

【請求項９】　プロセッサ間通信時にページフォールト

した論理ページ番号が逐次格納されたフォールトページリストを、送り元プロセッサがリモートリードを用いて読み出して、該フォールトページリストにある論理ページ番号のデータだけを再送することを特徴とする請求項８に記載のプロセッサ間通信装置。

【請求項１０】　前記再送の後で、前記パケットでページフォールトの有無を再度確認し、ページフォールトが起こらなくなるまで、再送をつづけることを特徴とする請求項９に記載のプロセッサ間通信装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、複数のプロセッサがネットワークによって結合された並列コンピュータでの、仮想記憶を実現するためのプロセッサ間通信装置に関する。

【０００２】

【従来の技術】複数のプロセッサが１つの問題を解く並列処理において、プロセッサ間通信は必須である。ここで、プロセッサ間通信は複数のプロセッサ間でのデータのやりとりをいう。また、このプロセッサ間通信は、複数のプロセッサで並列に処理することにより発生するものであって、並列処理のオーバヘッドの１つとなり、プロセッサ間通信を高速に行なうことが、並列処理効果を高めるためには必須となっている。

【０００３】一方、並列計算機を複数のユーザで使うには、１人のユーザが処理に使用するプロセッサを決めて、１つのプロセッサでは１人のユーザのジョブだけが走るように、システムを空間的に分割するスペースシェアリングという手法が採られてきた。しかしながら、この方法では、各ユーザが使用するプロセッサ数の合計は、物理的なプロセッサ数を越えられないという制限がある。このような制限を克服して、複数のユーザが並列計算機を使う方法の１つとして、時間的にプロセッサを分割して使うタイムシェアリングの方法がある。タイムシェアリングを行なうには、複数ユーザのジョブのメモリメージを外部の記憶媒体、例えばハードディスクに退避する仮想記憶のサポートが必須となる。

【０００４】また、科学技術計算では、プロセッサ性能の向上により、それが扱うデータ規模はますます増大していく。並列処理でのデータの分配は、まず、１つのプロセッサにディスクやホストマシンからデータを転送し、分割したデータをそれぞれのプロセッサに分配する方法が採られることが多い。このような場合、分配後にはデータは各プロセッサの主記憶に格納されるが、最初に１つのプロセッサの主記憶にデータをロードする時点で、データ量が主記憶容量を越えてしまうことがある。このような問題に対応するためにも、外部の記憶媒体、例えばハードディスクを使って実主記憶容量以上のデータを扱う仮想記憶のサポートが必須である。

【０００５】

【発明が解決しようとする課題】しかし、仮想記憶を並列計算機でサポートするには、次のような課題がある。即ち、プロセッサ間通信が高速化したため、ページアウトしたページをディスクから主記憶にページインする時間が相対的に非常に長くなったことである。そのため、ページインのためにプロセッサ間通信を止めた時の、他のプロセッサ間通信への影響の大きさが問題になってくる。

【０００６】図１１は、従来の並列計算機システムのページインの説明図で、（Ａ）は全ての領域がページインされている場合、（Ｂ）は通信バッファがページインされている場合、（Ｃ）はプロセッサ間通信により送信データと受信領域がページインされている場合を示している。従来は、図１１（Ａ）～（Ｃ）の構成により、、ページインのためにプロセッサ間通信を止めた時の、他のプロセッサ間通信への干渉を防止していた。

【０００７】図１１（Ａ）に示す従来方法１は、仮想記憶を使わないという解である。プロセッサ間通信の対象になる主記憶領域は限定できないので、プログラムが使用する全領域を実メモリに割り当てて、ページアウトしないことにするのである。この方法では、上に述べた並列計算機での仮想記憶サポートの要求に答えられないという課題がある。

【０００８】図１１（Ｂ）に示す従来方法２は、プロセッサ間通信のための送信バッファ、受信バッファを設けるものである。この方法では、送信バッファ、受信バッファだけを常に実メモリに割り付けておき、プロセッサ間通信するデータは必ず、この送信メモリバッファと受信メモリバッファを介して転送される。そうすることにより、プロセッサ間通信自体は、実メモリから実メモリへの転送となり、プロセッサ間通信を止めることはない。しかしながら、プロセッサ間通信には、送信バッファへのデータのコピーと、受信バッファから宛先アドレスへのデータのコピーの２回のデータコピーが必要となり、プロセッサ間通信性能が著しく低下するという課題がある。

【０００９】また、図１１（Ｃ）に示す従来方法３は、通信時にそのプロセッサ間通信の送信データの領域と受信領域を予め実メモリ上に割り付けてから、プロセッサ間通信を行なうものである。この方法では、コピーすることなくデータが宛先アドレスに転送されるが、最初に実メモリ上に割り当てられていることを確認するためのプロセッサ間のやりとりが必要となる。これにより、コピーを行なわなくても、実質的なプロセッサ間通信の性能が低下する。

【００１０】以上を簡潔にまとめると、従来方法１では仮想記憶自体をサポートできていない。従来方式２では、送信バッファへのデータのコピーと、受信バッファから宛先アドレスへのデータのコピーのオーバヘッドから、通信性能が著しく低下する。従来方式３では、たと

え、プロセッサ間通信に関わるメモリ領域が実メモリ領域に割り付けられていたとしても、必ず、プロセッサ間通信の前に実メモリへの割り付けを確認するためのプロセッサ間通信が必要である。そこで、この確認のためのプロセッサ間通信により、プロセッサ間通信の性能が低下する。即ち、従来の並列計算機システムで仮想記憶をサポートした場合に、実メモリ上に割り当てられた領域間のプロセッサ間通信の性能が低下してしまうという課題があった。

【００１１】本発明は上述する課題を解決するもので、プロセッサ間通信に関わるメモリ領域が実メモリ領域に割り付けられている場合でも、性能低下の原因になる作業をせずに、プロセッサ間通信の性能を良好に保持するプロセッサ間通信装置を提供することを目的とする。

【００１２】

【課題を解決するための手段】上記課題を解決する本発明の請求項１に記載のプロセッサ間通信装置は、図１に示すように、ＣＰＵ３、主記憶４、送信装置５、並びに受信装置６を有するプロセッサ１が複数個、ネットワーク２によって結合された並列コンピュータに用いられる。プロセッサ間通信装置は、プロセッサ間通信で通信されるデータの読み出しアドレスと書き込みアドレスが論理アドレスで指定され、送信装置５と受信装置６内に論理アドレスから物理アドレスに変換する手段ＴＬＢ１０、２１を有している。

【００１３】また、図１、図５に示すように、プロセッサ間通信装置は、タスクごとに割り当てられた通信ＩＤとプロセッサ間通信のパケットを送った送り元の論理プロセッサ番号から決まる主記憶上の位置２９に、該通信ＩＤを宛先とした該論理プロセッサ番号のプロセッサからのプロセッサ間通信によるデータ転送中に、ページフォールトが発生したかを示すフラグ２９１と、該データ転送中で直近にページフォールトした論理ページ番号２９２と、ページフォールトしたページを記憶するページリストの先頭アドレスを記憶する手段２９３と、ページフォールトしたページ数とを記憶する手段２９４を持つ構成である。

【００１４】好ましくは、本発明のプロセッサ間通信装置では、受信装置内でのアドレス変換の結果、ページの状態がページアウトの場合には、前記のページフォールトに関するデータを通信ＩＤと送り元論理プロセッサ番号によって決まる主記憶上の位置に記憶し、アドレス変換にかかる論理ページを主記憶上のフォールトページリスト３３に加える構成とするとよい。さらに好ましくは、図８のステップＳ１１０に示すように、アドレス変換にかかる論理ページを予め用意しておいたダミーページを割り当てる構成とするとよい。

【００１５】また、本発明の請求項４に記載のプロセッサ間通信装置では、図４に示すように、ＴＬＢ１０は、論理アドレス２０１を構成する論理ページ２０２の情報として、論理ページの割り当てられた物理ページ番号２２９と、ページの状態２３０として下記（ｉ）～（ｉｖ）の状態を持つことを特徴とするものである。

（ｉ）物理ページが割り当てられていない無効の状態；

（ｉｉ）現在主記憶になく外部記憶媒体に追い出されているページアウトの状態；

（ｉｉｉ）以前はページアウトの状態で現在ページイン処理を行なっている際中であるページイン中の状態；

（ｉｖ）該主記憶上に割り付けられているページインの状態。

【００１６】また、本発明の請求項５に記載のプロセッサ間通信装置は、図４に示すように、受信装置でのアドレス変換時に、タスクごとに割り当てられた通信ＩＤと、パケットの送り元の論理プロセッサ番号から決定される主記憶上の場所に記憶されている論理ページ番号と、変換しようとしている論理ページ番号とを比較する手段２２２、２２３を持つ構成としている。

【００１７】さらに、好ましくは、本発明のプロセッサ間通信装置では、図８に示すように、受信装置内でのアドレス変換の結果、ページの状態２３０がページイン中の場合には、比較器２２を用いて論理ページの比較を行なう。そして、論理ページ番号が同じ場合には何もせず、論理ページ番号が異なった場合にだけ、その論理ページ番号をフォールトページリスト３３に加え、直近にページフォールトした論理ページ番号２９２として記憶し、ページフォールトしたページ数２９４を１増やす、ことを行なう構成とするとよい。

【００１８】また、本発明の請求項７に記載のプロセッサ間通信装置では、図３に示すように、宛先の通信ＩＤと送り元の論理プロセッサ番号から決まる主記憶上の位置のデータと、パケットに付加されたデータとを交換し、交換した結果を送り元のプロセッサの決められた主記憶上の位置に書き込むチェックパケット１３３を持つ。

【００１９】好ましくは、本発明のプロセッサ間通信装置では、プロセッサ間通信でデータを送った直後に、チェックパケットを用いて、事前に送ったプロセッサ間通信の受信側でのページフォールトがあったかどうかを知る構成とするとよい。

【００２０】さらに、本発明のプロセッサ間通信装置では、第２の実施の形態で示すように、作成されるページフォールトしたページのリストを、リモートリードによって、送り元のプロセッサが読み出して、そこにある論理ページ番号のデータだけを再送し、再度前記パケットでページフォールトの有無を確認し、ページフォールトが起こらなくなるまで、再送をつづける構成としている。

【００２１】

【発明の実施の形態】以下、図面を用いて本発明の実施の形態を説明する。　図１は、本発明の一実施の形態を

説明する構成ブロック図である。図において、並列コンピュータは、複数のプロセッサ１がネットワーク２に接続されて構成されている。プロセッサ１は、ＣＰＵ３と主記憶４と送信装置５と受信装置６から構成される。送信装置５内は、ヘッダ読み出し回路７、ヘッダレジスタ８、パケット生成回路９、ＴＬＢ１０、割り込み発生回路１１、並びにＦＩＦＯメモリ１２から構成されている。

【００２２】ヘッダ読み出し回路７は、ＣＰＵ３からの送信要求によって、主記憶４上に設けられたパーセルヘッダキュー１３からヘッダ１４を読み出し、ヘッダの内容を検査し、エラーがなければ、ヘッダレジスタ８に格納する。ヘッダレジスタ８は、複数のヘッダ１４を格納できる。

【００２３】パケット生成回路９は、ヘッダレジスタ８内の複数のヘッダ１４のうち次にパケットを送信できるものを選び、そのヘッダに書かれたソースアドレスをＴＬＢ１０を使って論理アドレスから物理アドレスに変換する。そして、ヘッダレジスタ８内のヘッダ１４を使って、ネットワーク２に送り出すパケットのヘッダを作成し、ＴＬＢ１０により変換された物理ソースアドレスからのデータの読み出しを主記憶４に依頼し、読み出した送信データ１５をパケットとしてネットワーク２に送り出す。

【００２４】ＴＬＢ１０は、ヘッダレジスタ８内の通信ＩＤとソースアドレスから、論理アドレスを物理アドレスに変換する。ＴＬＢ１０内にはアドレス変換に必要な情報がキャッシュされているが、必要に応じて、主記憶４上のページテーブル１６にアクセスしたり、ＯＳにページの割り当てを要求するための割り込みをおこしたりして、アドレス変換に必要な情報を得る。

【００２５】ＦＩＦＯメモリ１２は、ネットワーク２とパケット生成回路９間に設けられたもので、データ幅の変換やエラー検出符号などの作成を行なう。割り込み発生回路１１は、ヘッダ読み出し回路７でのヘッダの検査の結果、エラーとなった場合や、ＴＬＢ１０における論理アドレスから物理アドレスへの変換の際に、ＯＳに対して処理を依頼しなければならない時に、ＣＰＵ３に対して割り込みを発生させる。

【００２６】受信装置６内は、ＦＩＦＯメモリ１７、通信ＩＤレジスタ１８、送り元論理プロセッサ番号レジスタ１９、書き込みアドレスレジスタ２０、ＴＬＢ２１、比較器２２、データ書き込み回路２３、データ交換回路２４、ヘッダ書き込み回路２５、並びに割り込み発生回路２６から構成される。

【００２７】ＦＩＦＯメモリ１７は、ネットワークから送られてきたパケットを一次的に格納する。通信ＩＤレジスタ１８は、パケットヘッダに含まれる通信ＩＤを格納する。送り元論理プロセッサ番号レジスタ１９は、パケットヘッダに含まれる、そのパケットを送信したプロセッサの論理プロセッサ番号を格納する。書き込みアドレスレジスタ２０は、パケットヘッダに含まれている、パケットのデータの書き込みアドレス（デスティネーションアドレス）を格納する。

【００２８】ＴＬＢ２１は、通信ＩＤレジスタ１８内の通信ＩＤと、書き込みアドレスレジスタ２０内のデスティネーションアドレスから、論理アドレスを物理アドレスに変換して、データ書き込み回路２３に物理書き込みアドレスを渡す。ＴＬＢ２１内にはアドレス変換に必要な情報がキャッシュされているが、必要に応じて、主記憶４上のページテーブル１６にアクセスしたり、ＯＳにページの割り当てを要求するための割り込みをおこしたりして、アドレス変換に必要な情報を得る。

【００２９】データ書き込み回路２３は、パケットで送られてきた受信データ２７を、主記憶４上の、ＴＬＢ２１から出力された書き込みアドレスに、書き込む。データ交換回路２４は、主記憶４上に設けられたチェックバッファ２８のエントリ２９（以下、チェックバッファエントリ２９という）の値と、パケットで送られてきたデータを交換し、チェックバッファエントリ２９の値をヘッダ書き込み回路２５に出力する。

【００３０】ヘッダ書き込み回路２５は、送られてきたパケットに含まれる情報と、データ交換回路２４からのデータとから、受信したパケットの返事のパーセルヘッダを作成し、主記憶４上の返事専用のパーセルヘッダキュー３４にヘッダを書き込み、送信装置５内のヘッダ読み出し回路７に送信要求を出す。

【００３１】比較器２２は、書き込みアドレスレジスタ２０に格納された書き込みアドレスの一部と、主記憶４上のチェックバッファエントリ２９の一部を比較し、比較結果をデータ書き込み回路２３に出力し、必要に応じて割り込み発生回路２６を使って割り込みを発生させる。

【００３２】割り込み発生回路２６は、比較器２２での比較結果で割り込みが必要な場合や、ＴＬＢ２１における論理アドレスから物理アドレスへの変換の際に、ＯＳに対して処理を依頼しなければならない時に、ＣＰＵ３に対して割り込みを発生させる。

【００３３】チェックバッファポインタテーブル３１は主記憶４に設けられたもので、各通信ＩＤのチェックバッファ２８へのポインタ３０を格納している。受信装置６は、通信ＩＤレジスタ１８内の通信ＩＤと、主記憶上のチェックバッファポインタテーブル３１を使って、その通信ＩＤに対応するチェックバッファ２８をアクセスする。

【００３４】宛先変換テーブル３２は、主記憶４に設けられたもので、送信装置５内のヘッダ読み出し回路７が、宛先論理プロセッサ番号を、宛先物理プロセッサ番号と通信ＩＤに変換するのに用いられる。フォールトページリスト３３は、主記憶４に設けられたもので、ペー

ジフォールトしたページの論理ページ番号を格納する。

【００３５】次に、この並列コンピュータで通信時に使われる、パーセルとパケットについて説明する。パーセルは通信の単位であって、そのヘッダ１４がパーセルヘッダキュー１３に書かれる。ヘッダレジスタ８には、ヘッダ読み出し回路７が、主記憶４上の宛先変換テーブル３２を使って、このパーセルヘッダの宛先を論理プロセッサ番号から、物理プロセッサ番号と通信ＩＤに、変換したものが書き込まれる。パケットはネットワーク内を流れる通信の単位であり、ネットワーク２を構成するスイッチ内のＦＩＦＯメモリの容量等により、その大きさが限定される。１つのパーセルで送られるデータはパケット生成回路９により、複数のパケットに分割され送られる。

【００３６】図２は、この並列コンピュータで通信されるリモートＤＭＡパーセルの構成図で、（Ａ）はパーセルヘッダキュー１３内での形式、（Ｂ）はヘッダレジスタ８内での形式、（Ｃ）はリモートＤＭＡパケットの形式を示している。

【００３７】まず、リモートＤＭＡパーセルは、送信データの読み出しアドレス１０５からデータ長１０４分のデータを、宛先論理プロセッサ番号１０３で示された宛先のプロセッサ１の主記憶４の書き込みアドレス１０６へ書き込む機能を持つ。リモートＤＭＡのパーセルヘッダ１０１の先頭にはコード欄１０２が設けられている。コード欄１０２に書き込まれた「０００」は、このパーセルヘッダがリモートＤＭＡのヘッダであることを示している。

【００３８】ヘッダ読み出し回路７により、主記憶４上のパーセルヘッダキュー１３から読み出されたリモートＤＭＡのパーセルヘッダ１０１は宛先変換テーブル３２を使って、宛先論理プロセッサ番号１０３が、宛先物理プロセッサ番号と通信ＩＤ１１０に変換されて、ヘッダレジスタ８に書き込まれる。

【００３９】パケット生成回路９は、ヘッダレジスタ８内のリモートＤＭＡのパーセルヘッダ１０８と、そのタスクが持っている情報１２１を使って、リモートＤＭＡのパケット１１４を生成する。リモートＤＭＡのパケット１１４は、リモートＤＭＡを示す「０００」が書き込まれたコード欄１１５と、宛先の物理プロセッサ番号と通信ＩＤ１１６と、パケット長１１７、送り元の論理プロセッサ番号１１８、書き込みアドレス１１９、送信データ１２０から構成される。

【００４０】パケット生成回路９は、ヘッダレジスタ８内の残りデータ長１１１と、送信データの読み出しアドレス１１２と書き込みアドレス１１３からパケット長１１７を決定し、リモートＤＭＡのパケット１１４を生成する。ヘッダレジスタ８内のリモートＤＭＡのパーセルヘッダの残りデータ長１１１と、読み出しアドレス１１２、書き込みアドレス１１３は、１つのパケットを送信するごとに更新され、パケット生成回路９は、残りデータ長１１１がゼロになるまで、パケットを送る。

【００４１】パケット生成回路９は、１つのパケットの送信や受信の途中にページフォールト等が起こらないように、読み出しアドレス１１２と書き込みアドレス１１３のページ境界をまたぐパケットは生成しない。即ち、パケット生成回路９は、読み出しアドレス１１２のページ境界と書き込みアドレス１１３のページ境界では、必ず、パケットが切られるように、パケット長１１７を決定する。

【００４２】次にチェックパケットのパーセルについて説明する。図３は、この並列コンピュータで通信されるチェックパケットのパーセルとチェックパケットの返事のパーセルの構成図で、（Ａ）はパーセルヘッダキュー１３内での形式、（Ｂ）はヘッダレジスタ８内での形式、（Ｃ）はチェックパケットの形式、（Ｄ）はパーセルヘッダキュー１３内での形式、（Ｅ）はヘッダレジスタ８内での形式、（Ｆ）はチェックパケットの返事のパケットの形式を示している。チェックパケットのパーセルは、特別なアドレスにある値と交換するデータ１２６を交換し、値をチェックパケットの返事のパケット１５２として返送し、指定された返事の書き込みアドレス１２５に書き込む機能を持つ。

【００４３】パーセルヘッダキュー内のチェックパケットのパーセルヘッダ１２２には、チェックパケットのパーセルであることを示す「１００」が書き込まれたコード欄１２３、宛先の論理プロセッサ番号１２４、チェックパケットの返事を書き込む書き込みアドレス１２５、交換するデータ１２６から構成される。チェックパケットのパーセルヘッダ１２２もヘッダレジスタ８に書き込まれる前に、宛先変換テーブル３２を使って、宛先論理プロセッサ番号１２４が宛先の物理プロセッサ番号と通信ＩＤ１３０に変換される。

【００４４】パケット生成回路９は、ヘッダレジスタ内のチェックパケットのパーセルヘッダ１２８と、そのタスクの情報１４１から、次のようなものから構成されるチェックパケット１３３を作成する。コード欄１３４に書き込まれた「１００」は、パケットの種類がチェックパケットであることを示す。送り元の論理プロセッサ番号１３７と送り元の物理プロセッサ番号と通信ＩＤ１３８は、そのタスクが持っている情報１４１である。返事の書き込みアドレス１３９と交換するデータ１４０は、チェックパケットのパーセルヘッダ１２８のものをそのまま使う。チェックパケットのパーセルヘッダ１２８は必ず１つのチェックパケット１３３として送信される。

【００４５】最後にチェックパケットの返事のパーセルヘッダ１４２について、図３を参照して説明する。このパーセルヘッダは受信装置６のデータ交換回路２４で交換されたチェックバッファエントリ２９と、チェックパケット１３３内にある情報とから作成され、ヘッダ書き

込み回路２５によって主記憶４上の返事専用のパーセルヘッダキュー３４に書き込まれる。

【００４６】チェックパケットの返事のパーセルヘッダ１４２には、チェックパケットの返事を示す「１０１」が書き込まれたコード欄１４３と、チェックパケット１３３の送り元の物理プロセッサ番号と通信ＩＤ１３８をそのまま使った宛先の物理プロセッサ番号と通信ＩＤ１４４、チェックパケット１３３の返事の書き込みアドレス１３９をそのまま使った返事の書き込みアドレス１４５と、交換したチェックバッファエントリ１４６から構成される。

【００４７】ヘッダ読み出し回路７は、返事専用のパーセルヘッダキューから、チェックパケットの返事のパーセルヘッダ１４２を読み出すと、チェックパケットの返事を示す「１０１」が書き込まれたコード欄１４３を確認する。他のリモートＤＭＡのパーセルヘッダ１０１やチェックパケットのパーセルヘッダ１２２の場合とは異なり、ヘッダ読み出し回路７は、宛先プロセッサ番号を変換しないで、パーセルヘッダキュー内のチェックパケットの返事のパーセルヘッダ１４２に書かれた宛先の物理プロセッサ番号と通信ＩＤ１４４をそのままヘッダレジスタ８に書き込む。これは、チェックパケットの返事のパーセルヘッダ１４２が、送信装置５のパケット生成回路９によって作られた送り元の物理プロセッサ番号と通信ＩＤ１３８をもとに、受信装置６のヘッダ書き込み回路２５によって、作られたものであり、ユーザタスクがアクセスできない特別なパーセルヘッダキュー３４に書かれたものであるため、そのまま使っても安全であるからである。

【００４８】パケット生成回路９は、チェックパケットの返事のパーセルヘッダ１４７からチェックパケットの返事のパケット１５２を作成し、ネットワークに送り出す。チェックパケットの返事のパーセル１４７も、必ず１つのチェックパケットの返事のパケット１４１となる。

【００４９】次にＴＬＢ１０、２１で行なわれるアドレス変換について図４を使って説明する。図４はＴＬＢ１０の詳細を説明する構成ブロック図である。なお、ＴＬＢ２１もＴＬＢ１０と同様の構成となっている。図４において、論理アドレス２０１は、論理ページ番号２０２とオフセット２０３によって構成されている。ＴＬＢ１０では、通信ＩＤ２０４と論理ページ番号２０２から、物理ページ番号２２９とページの状態２３０を引く。この実施例では、ＴＬＢ１０には４個のエントリがあり、与えられた通信ＩＤ２０４と論理ページ番号２０２と、４個のエントリ内の通信ＩＤ２０６～２０９と論理ページ番号２１０～２１３とが比較器２２２、２２３によって比較され、両方が同じものの物理ページ番号２１４～２１７とページの状態２１８～２２１が選択回路２２４からの信号により、ＭＵＸ２２５、２２６によって、選ばれ、そのエントリの物理ページ番号２２９とページの状態２３０が出力される。

【００５０】もし、４個のエントリ内に、通信ＩＤと論理ページ番号の組が与えられた通信ＩＤ２０４と論理ページ番号２０２に等しいものがなければ、ＴＬＢ１０は主記憶４上のページテーブル１６をアクセスして、該当するページ情報をＴＬＢ１０内の４個のエントリのうちの１個に読み出す。もし、ページテーブル１６にもそのページ情報がなければ、ＴＬＢ１０はＯＳに割り込みをあげて新しいページ情報を設定してもらう。

【００５１】主記憶４上のページテーブル１６の構成はいろいろなものが考えられる。最も簡単なものは、通信ＩＤ２０４と論理ページ番号２０２を組み合わせたものをオフセットとする構成である。主記憶４の使用効率等を考えると、論理アドレス２０１を論理ページ番号２０２とオフセット２０３だけではなく、セグメントと論理ページ番号とオフセットと解釈し、セグメントごとにページテーブル１６を用意する構成としてもよい。

【００５２】ただし、ページの状態２３０として、次のような状態を定義する必要がある。ページの状態２３０は２bitで表されており、次のようなものがある。

【００５３】“００”：そのページが無効であることを示す。新規にアクセスされたページの場合もあるので、割り込みを発生させて、ＯＳに物理ページの割り付けを依頼する。

【００５４】“０１”：そのページが主記憶４になくディスクにページアウトされている状態であることを示す。この場合も割り込みを発生させて、ＯＳに対し処理を依頼する。

【００５５】“１０”：そのページが主記憶４になくディスクにページアウトされているが、ページイン中であることを示す。この場合は、割り込みを発生させない。

【００５６】“１１”：そのページが主記憶４上にページインされていることを示す。

【００５７】次にチェックバッファエントリ２９について説明する。図５に示すように、チェックバッファエントリ２９には、ページフォールトが起こったかどうかを示すページフォールトフラグ２９１、直近にページフォールトした論理ページ番号２９２、ページフォールトしたページ数２９４、フォールトページリスト３３の先頭アドレス２９３が格納されている。ページフォールトフラグ２９１は、“０”の時にページフォールトが起こっていないことを、“１”の時に、ページフォールトがすでに起こったことを示す。

【００５８】次にフォールトページリスト３３について説明する。フォールドページリスト３３は、１つのチェックバッファエントリ２９に対して１つ設けられており、その先頭アドレス２９３はチェックバッファエントリ２９に書かれている。フォールトページリスト３３には、ページフォールトした論理ページ番号３３１が書き

込まれている。フォールトページリスト３３の大きさは、チェックバッファエントリ２９内のページフォールトしたページ数２９４となる。１つのパーセルヘッダで送信できるデータの大きさは限られているので、フォールトページリスト３３の大きさの上限は、「１パーセルで転送できるデータ」÷「１ページの大きさ」となる。

【００５９】次に本発明でのプロセッサ間通信の方法に関して説明する。まず、前述したように、パーセルが通信の単位である。１つのリモートＤＭＡパーセルでは限られた大きさの連続したアドレスのデータが通信できる。本実施例では、１つのリモートＤＭＡパーセルで送れるデータの大きさは５１２ｋバイトとする。また、１ページの大きさは４ｋバイトとする。従って、フォールトページリスト３３の大きさは最大で１２８ページ分となる。このくらいの大きさなら、フォールトページリスト３３の領域を仮想記憶のページアウトの対象から外し、必ずページインの状態に固定することも可能である。

【００６０】本発明では、１つのリモートＤＭＡパーセルを送るごとに、その直後にチェックパケットのパーセルを送る。

【００６１】図６に、通信ライブラリの処理を示す。まず、通信ライブラリは、送信するデータを１パーセルで送れる大きさにＮ分割する（ステップＳ２０１）。次に、Ｎが“０”かどうか調べ（ステップＳ２０２）、“０”でないなら、１個のリモートＤＭＡパーセルを送る（ステップＳ２０３）。その直後に１個のチェックパケットのパーセルを送る（ステップＳ２０４）。チェックパケットで送る交換するデータは、ページフォールトフラグ２９１が“０”のチェックバッファエントリ２９の初期値となる。

【００６２】そして、そのチェックパケットの返事を待つ（ステップＳ２０５）。チェックパケットの返事は、直前に送ったリモートＤＭＡパーセルのデータ受信時にページフォールトが起こったかどうかを示すチェックバッファエントリ２９である。そのチェックバッファエントリ２９のページフォールトフラグ２９１を調べる（ステップＳ２０６）。そして、もし、ページフォールトが起こっていれば、その処理を行なって、このリモートＤＭＡパーセルの送信を完了する（ステップＳ２０７）。

【００６３】ページフォールトに対する処理は、要するに、このリモートＤＭＡパーセルで送ったデータが正確に受信されるように処理するもので、詳細は後で説明する。一方、もし、ページフォールトが起こっていなければ、何もしない。そして、Ｎを１減らし（ステップＳ２０８）、Ｎがゼロになるまで、繰り返す。

【００６４】次に、図１、図２、及び図７を使って送信時のページフォールトに対する処理について説明する。送信装置５内のパケット生成回路９は、ヘッダレジスタ８内のパーセルヘッダを１つ選び（ステップＳ００１）、そのパケットを生成する。そのパーセルヘッダがリモートＤＭＡのパーセルヘッダである場合（ステップＳ００２）には、パケット生成回路９は、その送信データ読み出しアドレスをＴＬＢ１０を使って物理アドレスに変換する（ステップＳ００３）。そしてページの状態２３０を調べる（ステップＳ００４）。

【００６５】ページの状態２３０が無効“００”なら、ＴＬＢ１０は割り込みを発生させて（ステップＳ００５）、ＯＳに新しいページを割り当ててもらう（ステップＳ００６）。通常は転送するデータは事前にＣＰＵが作成しているので、読み出しアドレスのページの状態２３０が“００”であることはほとんどない。そして、ＯＳはページの状態２３０を“１１”にして（ステップＳ００７）、ＴＬＢ１０を割り込み状態から復帰する（ステップＳ００８）。ＴＬＢ１０は、また、ページの状態２３０を調べる（ステップＳ００４）。

【００６６】ページの状態２３０がページアウト“０１”なら、ＴＬＢ１０は割り込みを発生させて（ステップＳ００９）、ＯＳにページインの操作を開始してもらう（ステップＳ０１０）。さらにＯＳは、ページの状態２３０を“１０”に書き変えページイン中の状態にする（ステップＳ０１１）。そして、ＯＳはＴＬＢ１０を割り込み状態から復帰させる（ステップＳ０１２）。ＴＬＢ１０は、また、ページの状態２３０を調べる（ステップＳ００４）。

【００６７】ページの状態２３０がページイン中“１０”なら、ＴＬＢ１０は割り込みを発生させず、ページの状態２３０をパケット生成回路９に伝える。パケット生成回路９は、そのパケットのデータがページイン中であるので、そのパーセルヘッダのパケットの送信を中止し（ステップＳ０１３）、他のパーセルヘッダの送信を試みるため、パーセルヘッダを選ぶ（ステップＳ００１）。

【００６８】ページの状態２３０がページイン“１１”なら、ＴＬＢ１０は割り込みを発生させず、ページの状態２３０をパケット生成回路９に伝える。パケット生成回路９は、主記憶４の送信データ１５をアクセスし、このパケットを送信する（ステップＳ０１４）。そして、パケット長をもとに、ヘッダレジスタ内の残りデータ長、読み出しアドレス、書き込みアドレスを更新する（ステップＳ０１５）。そして、送ったパケットが最後かどうか、つまり、残りデータ長がゼロかどうかを調べる（ステップＳ０１６）。

【００６９】もし、最後のパケットでなければ、次のパケットを送るためパーセルヘッダの選択を行なう（ステップＳ００１）。もし、最後のパケットの場合には、ヘッダレジスタ８からこのパーセルヘッダを消して、ヘッダレジスタ８を解放する（ステップＳ０１７）。そして、次に送るパーセルヘッダを選択する（ステップＳ００１）。

15

【００７０】パケット生成回路９が選んだパーセルヘッダが、リモートＤＭＡのパーセルではなかった場合、つまり、チェックパケットのパーセルヘッダか、チェックパケットの返事のパーセルヘッダの場合には、ＴＬＢ１０を使ってアドレスの変換をする必要はなく、そのままパケットを送信する。これらのパーセルは１個のパケットにしかならないので、パーセルヘッダをヘッダレジスタ８から消し、ヘッダレジスタを解放する（ステップＳ０１７）。そして、次のパーセルヘッダを選択する（ステップＳ００１）。

【００７１】次に、図１、図５、及び図８を使って、受信側でのページフォールトに対する処理について説明する。受信装置６は、リモートＤＭＡのパケットが到着すると、パケットヘッダの情報を、通信ＩＤレジスタ１８、送り元プロセッサ番号レジスタ１９、書き込みアドレスレジスタ２０に格納する（ステップＳ１０１）。そして、通信ＩＤレジスタ１８内の通信ＩＤと、書き込みアドレスレジスタ２０内の書き込みアドレスから、ＴＬＢ２１を使って、論理アドレスを物理アドレスに変換する（ステップＳ１０２）。

【００７２】ＴＬＢ２１は、入力された通信ＩＤと論理ページ番号から、そのページの物理ページ番号とページの状態２３０を引く。そして、そのページの状態２３０を調べる（ステップＳ１０３）。

【００７３】ページの状態２３０が無効“００”なら、ＴＬＢ２１は割り込みを発生させて、ＯＳに新しいページの割り当てを行なってもらう（ステップＳ１０４）。ＯＳは、主記憶上にある新しいページをその論理ページに割り当てて物理ページ番号を設定し（ステップＳ１０５）、ページの状態２３０を“１１”にし（ステップＳ１０６）、割り込み処理から抜ける（ステップＳ１０７）。ＴＬＢ２１は、新たに設定されたページ情報を調べる（ステップＳ１０３）。

【００７４】ページの状態２３０がページアウト“０１”なら、ＴＬＢ２１は割り込みを発生させて、ＯＳに以下のような処理を依頼する（ステップＳ１０８）。まず、ＯＳはそのページのページインの処理を開始する（ステップＳ１０９）。そして、ＯＳは受信データを書き捨てるために予め用意していたダミーページをその物理ページ番号に設定し（ステップＳ１１０）、ページの状態２３０をページイン中“１０”にする（ステップＳ１１１）。

【００７５】次にＯＳは、このパケットの通信ＩＤと送り元の論理プロセッサ番号から、該当するチェックバッファエントリ２９をアクセスし、ページフォールトフラグ２９１を調べる（ステップＳ１２０）。もし、以前にページフォールトが起こっていない場合には、ページフォールトフラグ２９１を“１”にして、フォールトページリスト３３の先頭アドレスをチェックバッファエントリ２９に設定する（ステップＳ１２１）。

16

【００７６】そして、チェックバッファエントリ２９にこの論理ページ番号を、直近にページフォールトした論理ページ番号２９２として書き込み（ステップＳ１２２）、チェックバッファエントリ２９のページフォールトしたページ数を１増やし（ステップＳ１２３）、フォールトページリスト３３にこの論理ページ番号を加える（ステップＳ１２４）。そして、割り込み処理から復帰する（ステップＳ１２５）。ＴＬＢ２１は、新たに設定されたページ情報を調べる（ステップＳ１０３）。

【００７７】ページの状態２３０がページイン中“１０”なら、ＴＬＢ２１は割り込みを発生させず、ページの状態２３０を受信装置６に伝える。受信装置６は、そのパケットの通信ＩＤからチェックバッファポインタテーブル３１をアクセスし、その通信ＩＤのチェックバッファの先頭アドレス３０を取り出し、送り元論理プロセッサ番号レジスタ１９内のそのパケットの送り元の論理プロセッサ番号をオフセットとして、チェックバッファエントリ２９をアクセスする。そして、比較器２２を使って、そこに書かれた論理ページ番号と、このページの論理ページ番号とを、比較する（ステップＳ１１３）。

【００７８】比較の結果、２つのページ番号が同じならば、データ書き込み回路２３は、その物理アドレスに受信データを書き込む。この物理アドレスはダミーのページであり、送られてきたデータは捨てられることになる（ステップＳ１１７）。比較の結果、２つのページ番号が異なれば、比較器２２は割り込みを発生させて（ステップＳ１１５）、ＯＳに次のことを依頼する。

【００７９】ＯＳは、このパケットの通信ＩＤと送り元の論理プロセッサ番号から、該当するチェックバッファエントリ２９をアクセスし、ページフォールトフラグ２９１を調べる（ステップＳ１２０）。もし、以前にページフォールトが起こっていない場合には、ページフォールトフラグ２９１を“１”にして、フォールトページリスト３３の先頭アドレスをチェックバッファエントリ２９に設定する（ステップＳ１２１）。

【００８０】そして、チェックバッファエントリ２９にこの論理ページ番号を、直近にページフォールトした論理ページ番号２９２として書き込み（ステップＳ１２２）、チェックバッファエントリ２９のページフォールトしたページ数を１増やし（ステップＳ１２３）、フォールトページリスト３３にこの論理ページ番号を加える（ステップＳ１２４）。そして、割り込み処理から復帰する（ステップＳ１２５）。そして、データ書き込み回路２３は、その物理アドレスに受信データを書き込む。この物理アドレスはダミーのページであり、送られてきたデータは捨てられることになる（ステップＳ１１７）。

【００８１】ページの状態２３０がページイン“１１”なら、ＴＬＢ２１は割り込みを発生させず、ページ情報をデータ書き込み回路２３に伝える。データ書き込み回

路２３は、主記憶にアクセスして、送られてきたデータを書き込む（ステップＳ１１８）。このパケットは正常に受信されたことになる。

【００８２】最後に、受信時にページフォールトが起こった場合の送信側での通信ライブラリの処理について説明する。チェックパケットのパーセルを送って、その返事を調べた時、ページフォールトフラグ２９１が“１”であった場合には、ページフォールトが起こったということである。通信ライブラリは、割り込みをおこして、ＯＳに対し、ページフォールトしたページへのデータの転送を依頼し、スリープする。

【００８３】ＯＳは、受信側のＯＳと通信を行ない、チェックバッファエンドリ２９に書かれたページフォールトしたページ数と、フォールトページリスト３３の先頭アドレスを使って、ダミーページに書き込まれて捨てられたデータを再送する。そして、リモートＤＭＡのパーセルで送ろうとしたデータが全部受信された時点で、スリープしていた通信ライブラリをおこす。

【００８４】第２の実施の形態では、新たにリモートリードパーセルを導入する。図９は、リモートリードパーセルとチェックパケットの返事のパーセルの構成図で、（Ａ）はパーセルヘッダキュー１３内での形式、（Ｂ）はヘッダレジスタ８内での形式、（Ｃ）はリモートリードパケットの形式、（Ｄ）はパーセルヘッダキュー１３内での形式、（Ｅ）はヘッダレジスタ８内での形式、（Ｆ）はリモートリードの返事のパケットの形式を示している。

【００８５】リモートリードパーセルは、図９に示すように、宛先論理プロセッサ番号４０３で指定されたプロセッサの読み出しアドレス４０５から、データ長４０４分のデータを読み出し、リモートリードパーセルを送ったプロセッサの主記憶４の書き込みアドレス４０６に書き込む機能を持つ。コード欄４０２に書き込まれた「１１０」は、これがリモートリードのパーセルヘッダ４０１であることを示す。リモートリードのパーセルヘッダ４０１もヘッダレジスタ８に書き込まれる前に、宛先変換テーブル３２を使って、宛先論理プロセッサ番号４０３が宛先の物理プロセッサ番号と通信ＩＤ４１０に変換される。

【００８６】パケット生成回路９は、ヘッダレジスタ内のリモートリードのパーセルヘッダ４０８と、そのタスクの情報４２２から、次のようなものから構成されるリモートリードパケット４１４を作成する。「１１０」が書き込まれたコード欄４１５は、それがリモートリードのパケットであることを示す。送り元の物理プロセッサ番号と通信ＩＤ４１８は、そのタスクが持っている情報１４１である。データ長４１９と読み出しアドレス４２０、書き込みアドレス４２１は、リモートリードのパーセルヘッダ４０８のものをそのまま使う。リモートリードのパーセルヘッダ４０８は必ず１つのリモートリードのパケット４１４として送信される。

【００８７】次に、リモートリードの返事のパーセルヘッダ４２３について説明する。このパーセルヘッダは受信装置６のヘッダ書き込み回路２５によって、リモートリードパケット４１４内にある情報から作成され、主記憶４上の返事専用のパーセルヘッダキュー３４に書き込まれる。

【００８８】リモートリードの返事のパーセルヘッダ４２３には、リモートリードの返事を示す「１１１」が書き込まれたコード欄４２４と、リモートリードパケット４１４の送り元の物理プロセッサ番号と通信ＩＤ４１８をそのまま使った宛先の物理プロセッサ番号と通信ＩＤ４２５、データ長４２６、読み出しアドレス４２７、書き込みアドレス４２８から構成される。

【００８９】ヘッダ読み出し回路７は、返事専用のパーセルヘッダキュー３４から、リモートリードの返事のパーセルヘッダ４２３を読み出すと、リモートリードの返事を示す「１１１」が書き込まれたコード欄４２４を確認し、チェックパケットの返事のパーセルヘッダの場合と同じように、宛先プロセッサ番号を変換しないで、パーセルヘッダキュー内のリモートリードの返事のパーセルヘッダ４２３に書かれた宛先の物理プロセッサ番号と通信ＩＤ４２５をそのままヘッダレジスタ８に書き込む。

【００９０】これは、リモートリードの返事のパーセルヘッダ１４２が、送信装置のパケット生成回路９によって作られた送り元の物理プロセッサ番号と通信ＩＤ４１８をもとに、受信装置６のヘッダ書き込み回路２５によって、作られたものであり、ユーザタスクがアクセスできない特別なパーセルヘッダキュー３４に書かれたものであるため、そのまま使っても安全であるからである。

【００９１】パケット生成回路９は、リモートＤＭＡのパーセルを処理するのと同じように、残りデータ長４３２と、読み出しアドレス４３３、書き込みアドレス４３４から、パケット長４３８を決定し、リモートリードの返事のパケット４３５を送り出す。リモートリードの返事のパケットは、複数送られることがある。

【００９２】リモートリードの返事のパーセルはリモートＤＭＡのパーセルと同じように、送信装置５で処理される。つまり、送信アドレスをＴＬＢ１０で論理アドレスから物理アドレスに変換しないといけないので、送信データがページアウトしている可能性もある。リモートリードの返事のパーセルは、図７でのリモートＤＭＡのパーセルの送信処理と同じく処理される。従って、リモートリードの返事のパーセルを送る時の送信側のページフォールトの問題は適切に処理される。

【００９３】第２の実施の形態では、このリモートリードパーセルを使って、フォールトページリスト３３を送信側通信ライブラリが読み、該当するページの部分を再送するものである。

【００９４】図１０は通信ライブラリの処理を説明する図で、ここでは適宜図５も参照して説明する。通信ライブラリは、１個のリモートＤＭＡのパーセルを送った（ステップＳ３０１）後に、チェックパケットを送る（ステップＳ３０２）。そして、そのチェックパケットの返事を受信（ステップＳ３０３）し、チェックバッファエントリ２９のページフォールトフラグ２９１を調べる（ステップＳ３０４）。もし、ページフォールトが起こっていなければ、次のパーセルの送信を行なう。

【００９５】もし、ページフォールトが発生していたら、送られてきたチェックバッファエントリ２９内にある、フォールトページリスト３３の先頭アドレスから、ページフォールトしたページ分のデータをリモートリードパーセルによって読み出す（ステップＳ３０５）。そして、フォールトページリスト３３に書かれた論理ページに送ったデータを再送する（ステップＳ３０６）。

【００９６】この時、ページフォールトしたページは連続していないかもしれないが、フォールトページリスト３３に書かれている順番、つまり、論理ページ番号が大きくなる順番に、そのデータを複数のパーセルで送る。そして、最後に、また、チェックパケットを送信してページフォールトの有無を確認する（ステップＳ３０２）。

【００９７】通信ライブラリは、ページフォールトしたページがすでにページインされていることを期待して再送するが、まだ、ページインされていないかもしれない。もし、ページインされていないページがあれば、また、受信側でダミーページに書き捨てられることになる。通信ライブラリは、ページフォールトしたページ分だけのデータを再送し、ページフォールトしたページはいつかはページインされるので、いつかは、再送でのページフォールトが発生しなくなる。

【００９８】第２の実施の形態によると、ＯＳを介さずに、ユーザタスクが再送をする構成としているので、ＯＳの負担を軽減できるという効果がある。また、フォールトページリスト３３に関しては、前述したように、あまり大きなリストにはならないので、事前にページイン状態に固定した領域を使うことで、フォールトページリスト３３そのもののページフォールトを避けることができる。

【００９９】さらに、たとえフォールトページリスト３３がページアウトされていても、リモートリードの返事のパーセルは、図７で示したリモートＤＭＡのパーセルと同じく処理されるので、送信データに関するページフォールトは適切に処理される。一方、リモートリードの返事のパケットの受信時のページフォールトであるが、フォールトページリスト３３の大きさは、たかだか１ページ分であることと、リモートリードしている側が受信側なので、受信するページをＣＰＵがアクセスすることで、事前にページインできることから、ページフォールトは回避できると考えられる。

【０１００】

【発明の効果】以上説明したように、請求項１に記載のプロセッサ間通信装置によれば、プロセッサ間通信に使われる領域の大きさを意識しないでプログラミングできることである。なぜなら、チェックバッファエントリとフォールトページリストは主記憶に記憶されるが、プログラムで使われる領域はすべて仮想記憶の対象なので、従来方式３のようにプロセッサ間通信する主記憶の領域を制限する必要がないからである。

【０１０１】請求項３に記載のプロセッサ間通信装置によれば、ページフォールトが発生しても、ネットワークの閉塞が起こらないという効果がある。なぜなら、ページフォールトしたページへのデータの書き込みは、通常のデータ受信動作と変わらず、ダミーページに書き込むことで、捨てられるからである。

【０１０２】請求項４に記載のプロセッサ間通信装置によれば、プロセッサ間通信の性能を低下させずにすむことである。なぜなら、論理ページの状態として、無効、ページアウト、ページイン中、及びページインの４状態に分けて、ページフォールトの確認を円滑に行なうからである。

【０１０３】請求項５に記載のプロセッサ間通信装置によれば、ページフォールトが発生した場合でも受信装置が発生する割り込みが少ないことである。なぜなら、チェックバッファエントリ内の直近にページフォールトした論理ページ番号と、今ページフォールトした論理ページ番号を受信装置が比較するため、同じページのページフォールトによる複数の割り込みを発生させずにすむからである。

【０１０４】請求項７及び請求項８に記載のプロセッサ間通信装置によれば、ページフォールトが発生しない場合には、プロセッサ間通信の性能を低下させずにすむことである。なぜなら、ページフォールトの確認は大量のデータを送った後に１度だけ、しかも、ページフォールトしたことを確認するためのパケットの処理を受信装置が高速に行なうからである。

【０１０５】請求項９に記載のプロセッサ間通信装置によれば、ページフォールトした時にＯＳの負担を軽減できるという効果がある。なぜなら、第２の実施の形態で導入したリモートリードパーセルを使うことにより、ユーザタスクがフォールトページリストを読み出して、自分でページフォールトで受信されなかったデータを再送することができるからである。

【図面の簡単な説明】

【図１】　本発明のプロセッサ間通信装置の構成図である。

【図２】　この並列コンピュータで通信されるリモートＤＭＡパーセルの構成図である。

【図３】　この並列コンピュータで通信されるチェック

パケットのパーセルとチェックパケットの返事のパーセルの構成図である。

【図４】　本発明の実施の形態で用いるＴＬＢの構成図である。

【図５】　本発明で用いるチェックバッファエントリとフォールトページリストの構成図である。

【図６】　本発明での送信側の通信ライブラリの処理を説明するフローチャートである。

【図７】　本発明での送信側の読み出しアドレスがページフォールトした場合の送信装置での処理について説明するフローチャートである。

【図８】　本発明での受信側の書き込みアドレスがページフォールトした場合の受信装置での処理について説明するフローチャートである。

【図９】　本発明の第２の実施の形態で導入するリモートリードのパーセルとパケットの構成図である。

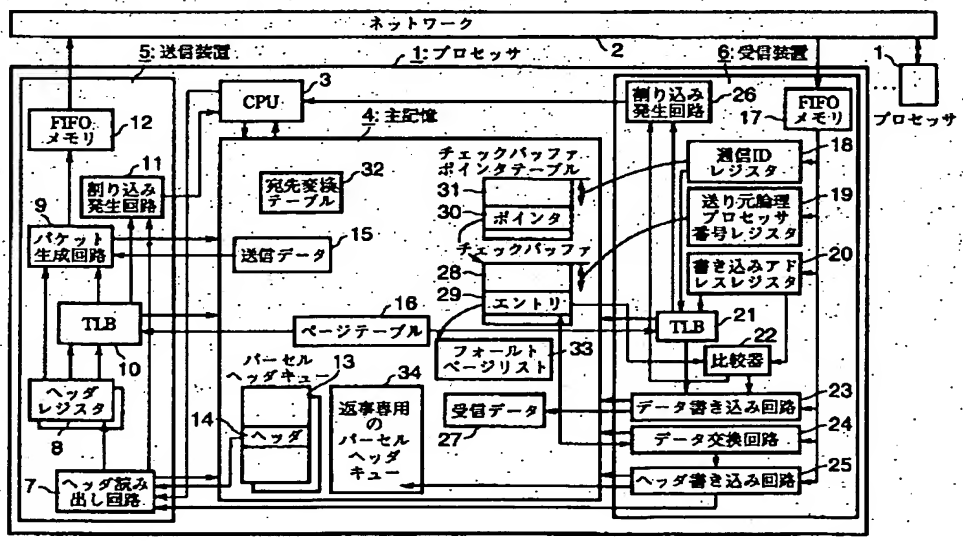
【図１０】　本発明の第２の実施の形態での、受信側でページフォールトが起こった場合の、通信ライブラリの処理について説明するフローチャートである。

【図１１】　従来の並列計算機システムのページインの説明図である。

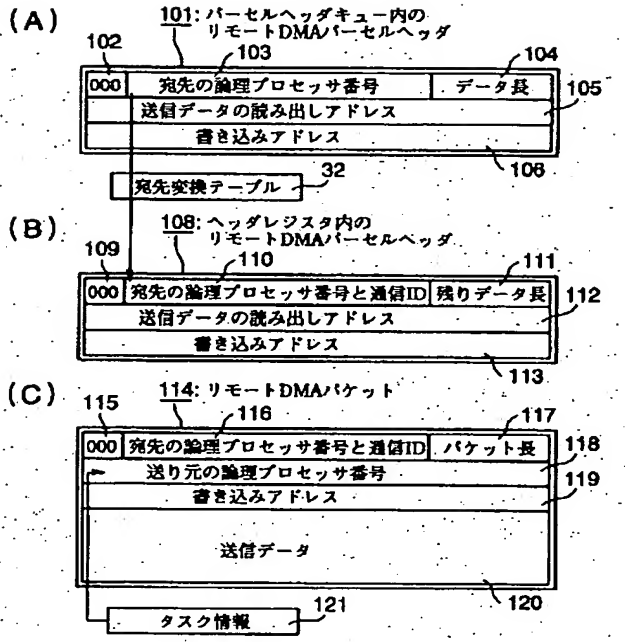
【符号の説明】

１　プロセッサ
２　ネットワーク
３　ＣＰＵ
４　主記憶
５　送信装置
６　受信装置
７　ヘッダ読み出し回路
＊８　ヘッダレジスタ
９　パケット生成回路
１０、２１　ＴＬＢ
１１　割り込み発生回路
１２　ＦＩＦＯメモリ
１３　パーセルヘッダキュー
１４　ヘッダ
１５　送信データ
１６　ページテーブル
１７　ＦＩＦＯメモリ
１８　通信ＩＤレジスタ
１９　送り元論理プロセッサ番号レジスタ
２０　書き込みアドレスレジスタ
２２　比較器
２３　データ書き込み回路
２４　データ交換回路
２５　ヘッダ書き込み回路
２６　割り込み発生回路
２７　受信データ
２８　チェックバッファ
２９　チェックバッファエントリ
３１　チェックバッファポインタテーブル
３２　宛先変換テーブル
３３　フォールトページリスト
１１４　リモートＤＭＡパケット
１３３　チェックパケット
１５２　チェックパケットの返事のパケット
４１４　リモートリードパケット
＊　４３５　リモートリードの返事のパケット

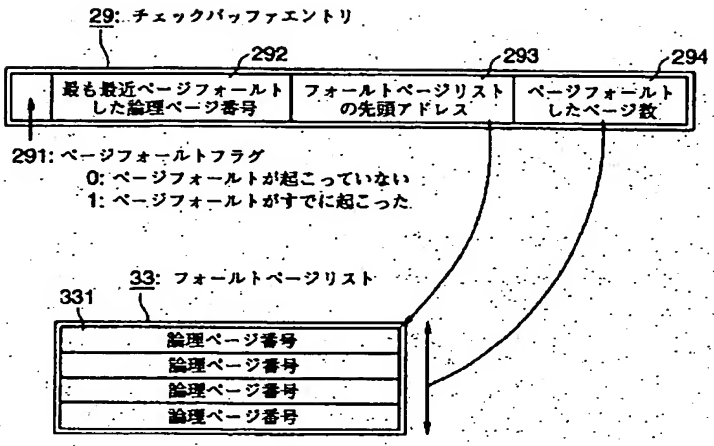
【図１】



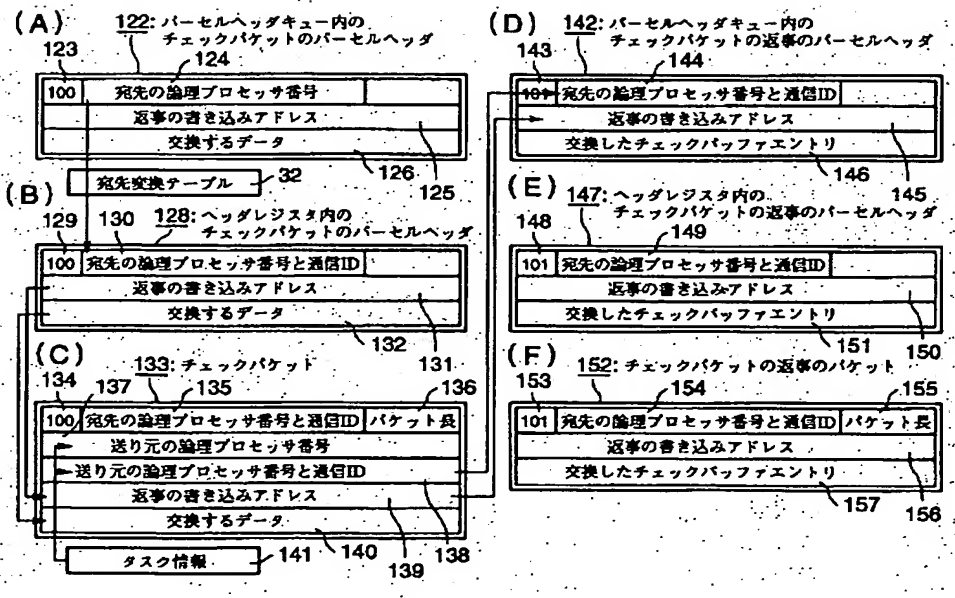
【図２】


(A)
101: パーセルヘッダキュー内の リモートDMAパーセルヘッダ
102
103
104
000
宛先の論理プロセッサ番号
データ長
105
送信データの読み出しアドレス
書き込みアドレス
106
宛先変換テーブル
32
(B)
108: ヘッダレジスタ内の リモートDMAパーセルヘッダ
109
110
111
000
宛先の論理プロセッサ番号と通信ID
残りデータ長
112
送信データの読み出しアドレス
書き込みアドレス
113
(C)
114: リモートDMAパケット
115
116
117
000
宛先の論理プロセッサ番号と通信ID
パケット長
118
送り元の論理プロセッサ番号
119
書き込みアドレス
送信データ
120
タスク情報
121

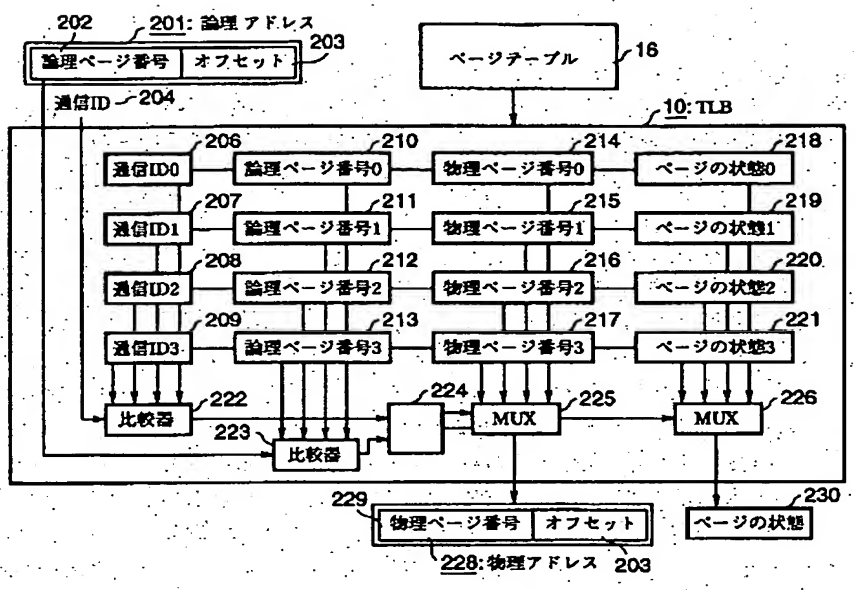
【図５】


29: チェックバッファエントリ
292
293
294
最も最近ページフォールトした論理ページ番号
フォールトページリストの先頭アドレス
ページフォールトしたページ数
291: ページフォールトフラグ
0: ページフォールトが起こっていない
1: ページフォールトがすでに起こった
33: フォールトページリスト
331
論理ページ番号
論理ページ番号
論理ページ番号
論理ページ番号

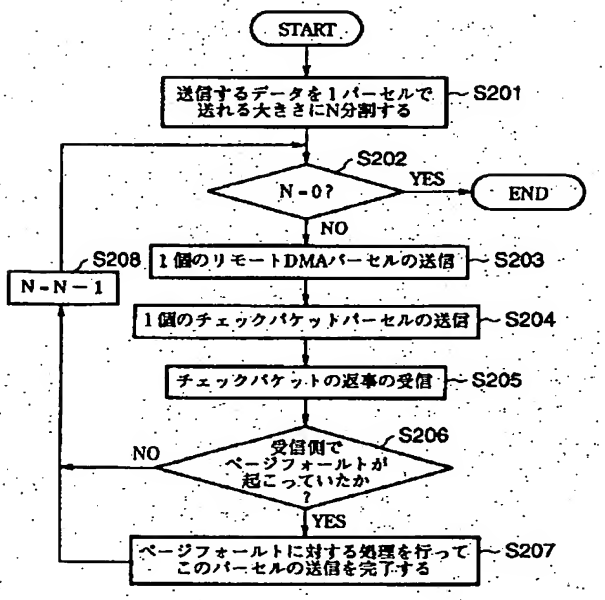
【図３】


(A)
122: パーセルヘッダキュー内の チェックパケットのパーセルヘッダ
123
124
100
宛先の論理プロセッサ番号
返事の書き込みアドレス
交換するデータ
126
125
宛先変換テーブル
32
(B)
129
130
128: ヘッダレジスタ内の チェックパケットのパーセルヘッダ
100
宛先の論理プロセッサ番号と通信ID
返事の書き込みアドレス
交換するデータ
132
131
(C)
133: チェックパケット
134
137
135
136
100
宛先の論理プロセッサ番号と通信ID
パケット長
送り元の論理プロセッサ番号
送り元の論理プロセッサ番号と通信ID
返事の書き込みアドレス
交換するデータ
タスク情報
141
140
139
138
(D)
142: パーセルヘッダキュー内の チェックパケットの返事のパーセルヘッダ
143
144
101
宛先の論理プロセッサ番号と通信ID
返事の書き込みアドレス
交換したチェックバッファエントリ
146
145
(E)
147: ヘッダレジスタ内の チェックパケットの返事のパーセルヘッダ
148
149
101
宛先の論理プロセッサ番号と通信ID
返事の書き込みアドレス
交換したチェックバッファエントリ
151
150
(F)
152: チェックパケットの返事のパケット
153
154
155
101
宛先の論理プロセッサ番号と通信ID
パケット長
返事の書き込みアドレス
交換したチェックバッファエントリ
157
156

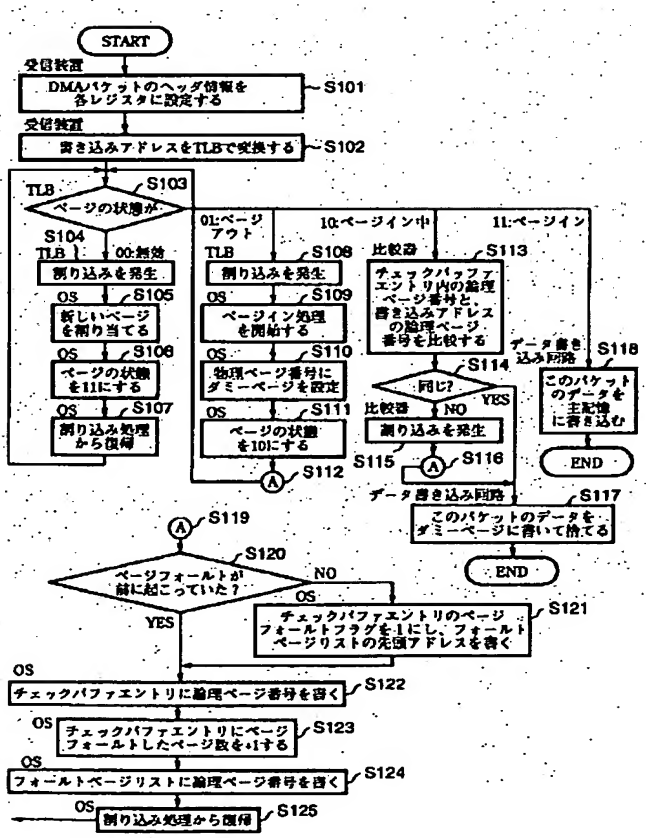
【図４】


202
201: 論理アドレス
論理ページ番号
オフセット
203
ページテーブル
16
通信ID
204
10: TLB
206
210
214
218
通信ID0
論理ページ番号0
物理ページ番号0
ページの状態0
207
211
215
219
通信ID1
論理ページ番号1
物理ページ番号1
ページの状態1
208
212
216
220
通信ID2
論理ページ番号2
物理ページ番号2
ページの状態2
209
213
217
221
通信ID3
論理ページ番号3
物理ページ番号3
ページの状態3
222
224
225
226
比較器
MUX
MUX
223
比較器
229
230
物理ページ番号
オフセット
ページの状態
228: 物理アドレス
203

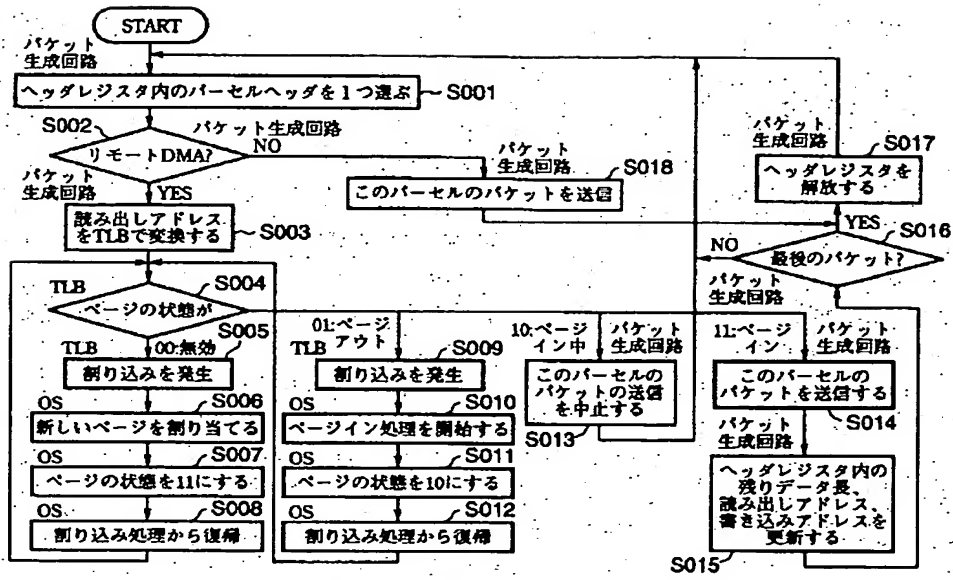
【図６】


START
送信するデータを１パーセルで送れる大きさにN分割する
S201
S202
N＝0?
YES
END
NO
S208
N＝N－１
１個のリモートDMAパーセルの送信
S203
１個のチェックパケットパーセルの送信
S204
チェックパケットの返事の受信
S205
S206
NO
受信側でページフォールトが起こっていたか?
YES
ページフォールトに対する処理を行ってこのパーセルの送信を完了する
S207

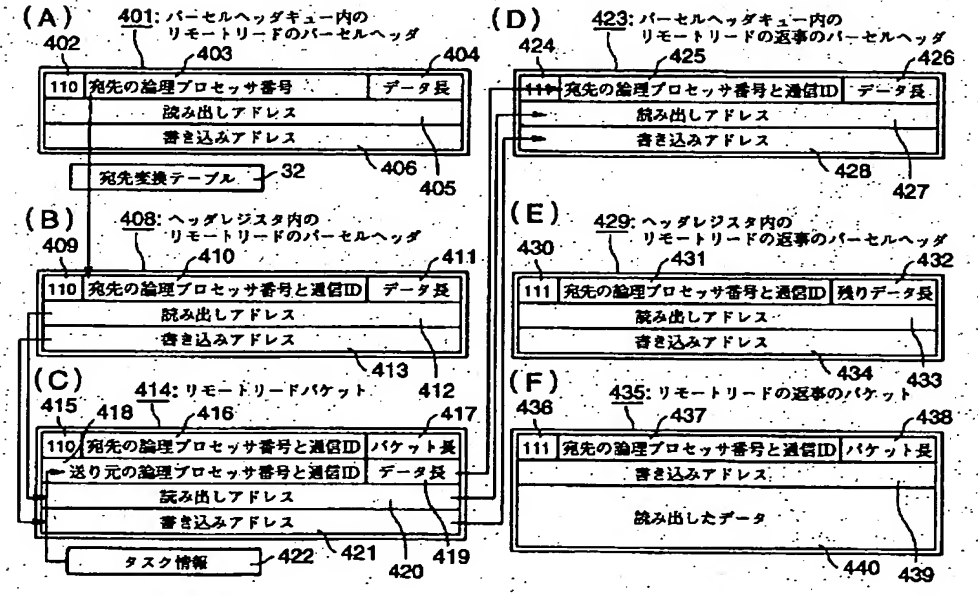
【図８】


START
受信装置
DMAパケットのヘッダ情報を各レジスタに設定する
S101
受信装置
書き込みアドレスをTLBで変換する
S102
TLB
S103
ページの状態が
01:ページアウト
10:ページイン中
11:ページイン
S104
TLB
00:無効
割り込みを発生
OS
S105
新しいページを割り当てる
OS
S106
ページの状態を11にする
OS
S107
割り込み処理から復帰
TLB
S108
割り込みを発生
OS
S109
ページイン処理を開始する
OS
S110
物理ページ番号にダミーページを設定
OS
S111
ページの状態を10にする
A
S112
比較器
S113
チェックバッファエントリ内の論理ページ番号と、書き込みアドレスの論理ページ番号を比較する
S114
同じ?
YES
NO
比較器
割り込みを発生
S115
A
S116
データ書き込み回路
S117
このパケットのデータをダミーページに書いて捨てる
END
データ書き込み回路
S118
このパケットのデータを主記憶に書き込む
END
A
S119
S120
ページフォールトが前に起こっていた?
NO
YES
OS
チェックバッファエントリのページフォールトフラグを１にし、フォールトページリストの先頭アドレスを書く
S121
OS
チェックバッファエントリに論理ページ番号を書く
S122
OS
チェックバッファエントリにページフォールトしたページ数を+1する
S123
OS
フォールトページリストに論理ページ番号を書く
S124
OS
割り込み処理から復帰
S125

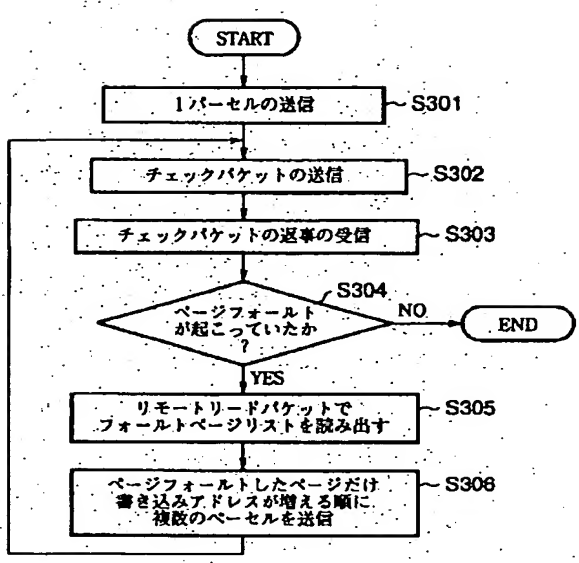
【図７】


START
パケット生成回路
ヘッダレジスタ内のパーセルヘッダを１つ選ぶ
S001
S002
リモートDMA?
パケット生成回路
NO
YES
パケット生成回路
読み出しアドレスをTLBで変換する
S003
パケット生成回路
このパーセルのパケットを送信
S018
パケット生成回路
ヘッダレジスタを解放する
S017
YES
S016
NO
最後のパケット?
パケット生成回路
TLB
S004
ページの状態が
TLB
00:無効
S005
割り込みを発生
OS
S006
新しいページを割り当てる
OS
S007
ページの状態を11にする
OS
S008
割り込み処理から復帰
01:ページアウト
TLB
S009
割り込みを発生
OS
S010
ページイン処理を開始する
OS
S011
ページの状態を10にする
OS
S012
割り込み処理から復帰
10:ページイン中
パケット生成回路
このパーセルのパケットの送信を中止する
S013
11:ページイン
パケット生成回路
このパーセルのパケットを送信する
S014
パケット生成回路
ヘッダレジスタ内の残りデータ長、読み出しアドレス、書き込みアドレスを更新する
S015

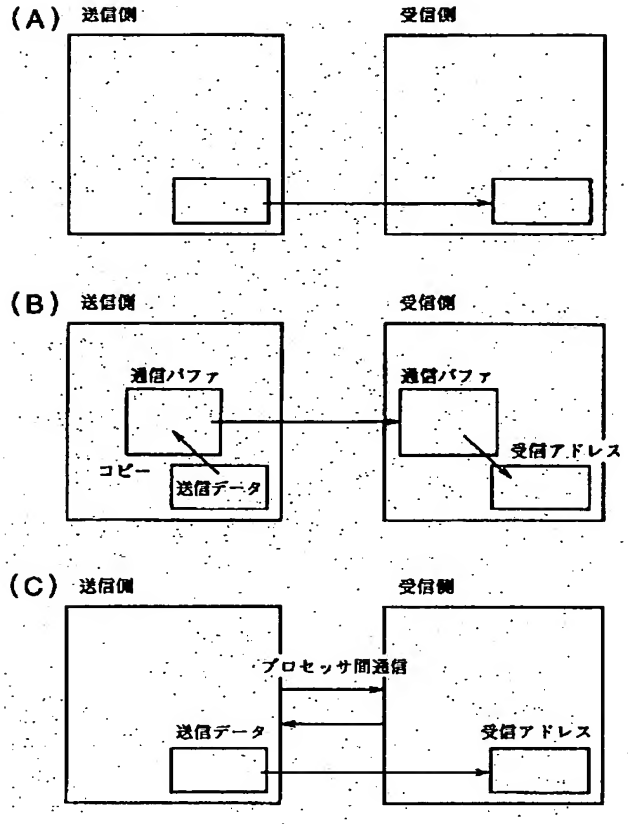
【図９】


(A)
401: パーセルヘッダキュー内のリモートリードのパーセルヘッダ
402
403
404
110
宛先の論理プロセッサ番号
データ長
読み出しアドレス
書き込みアドレス
406
405
宛先変換テーブル
32
(B)
408: ヘッダレジスタ内のリモートリードのパーセルヘッダ
409
410
411
110
宛先の論理プロセッサ番号と通信ID
データ長
読み出しアドレス
書き込みアドレス
413
412
(C)
414: リモートリードパケット
415
418
416
417
110
宛先の論理プロセッサ番号と通信ID
パケット長
送り元の論理プロセッサ番号と通信ID
データ長
読み出しアドレス
書き込みアドレス
タスク情報
422
421
420
419
(D)
423: パーセルヘッダキュー内のリモートリードの返事のパーセルヘッダ
424
425
426
111
宛先の論理プロセッサ番号と通信ID
データ長
読み出しアドレス
書き込みアドレス
428
427
(E)
429: ヘッダレジスタ内のリモートリードの返事のパーセルヘッダ
430
431
432
111
宛先の論理プロセッサ番号と通信ID
残りデータ長
読み出しアドレス
書き込みアドレス
434
433
(F)
435: リモートリードの返事のパケット
436
437
438
111
宛先の論理プロセッサ番号と通信ID
パケット長
書き込みアドレス
読み出したデータ
440
439

【図１０】


START
1パーセルの送信
S301
チェックパケットの送信
S302
チェックパケットの返事の受信
S303
S304
ページフォールトが起こっていたか？
NO
END
YES
リモートリードパケットでフォールトページリストを読み出す
S305
ページフォールトしたページだけ書き込みアドレスが増える順に複数のパーセルを送信
S306

【図１１】


(A)
送信側
受信側
(B)
送信側
受信側
通信バファ
通信バファ
コピー
送信データ
受信アドレス
(C)
送信側
受信側
プロセッサ間通信
送信データ
受信アドレス